Integration of Molecular Features with Clinical Information for Predicting Outcome for Neuroblastoma Patients

Yatong Han,[1] Jie Zhang,[2] Chao Wang,[3] Xiufen Ye,[1*] Yusong Liu,[1] Kun Huang[2*]

[1]Department of Automation, Harbin Engineering University, Harbin, China.
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio 43210, USA.
[3]Senior software engineer, Thermo Fisher Scientific
[*]Corresponding Authors

Abstract

Neuroblastoma (NB) is the most common extracranial solid tumor in children. NB in about 50% of pediatric patients will metastasize and result in a poor outcome. In order to provide better prognosis and facilitate individualized precise treatment, here we developed a novel workflow, which integrates clinical information and molecular features such as gene expression for prognosis. First, we mined co-expressed gene modules from microarray and RNA-seq data using the weighted network mining algorithm lmQCM; secondly, we build weight matrix with module eigengenes and a consensus clustering method called Molecular Regularized Consensus Patient Stratification (MRCPS), which aggregates both essential clinical information and multiple eigengene data for patient stratification. Our method improves prognosis significantly by regularizing clinical partition of patients using the additional weight matrix information. Our results suggested this method has a superior performance for predicting survival than only use genetic data and clinical diagnose result. Simultaneously, a subgroup of patients with extremely poor survival in early months was identified.

Keyword: neuroblastom survival time predict; gene co-expression network; consensus cluster

1 Introduction

Neuroblastoma (NB) is one of the most common cancers in children. About 50% of pediatric patients with NB will suffer metastasis and have a poor outcome. Accurate prognosis of patients will help to establish a individualized precise treatment plan for the patients, and lead to an improved long-term survival rates. Currently, clinical stage and risk at diagnosis are strong prognostic factors for NB[1]. However, with the accumulation of genomic and pathological data, an ideal approaches to address improve the prediction accuracy is to integrate genetic mutations, gene expression profiles, tissue and organ morphological features as well as clinical phenotypes to make a holistic decision. To achieve this goal, new methods for data integration are needed. Specifically, consensus clustering method, which integrate multiple clustering results from different types of data to achieve a single clustering of the data, has been introduced for this purpose. Currently there are two major approaches to achieve the consensus learning goal: 1) Probabilistic approach, which adopts a maximum likelihood formulation to generate the consensus clustering results using the Dirichlet mixture model given the distributions of base labels[2]; and 2) Similarity approach, which directly finds consensus clusters that agree the most with the input base clusters[3]. However, most of the consensus learning algorithms cannot be directly applied to multi-modal data with mixed data types (e.g., numerical data of transcriptomic levels of genes and categorical data for clinical stages of the patients). This limits the clinical applications of consensus learning algorithms. In this work, we present an effective integration workflow for numeric transcriptomic data and categorical clinical information. *Our goal is to find a consensus partition of patients from transcriptomic data and clinical attributes in order to reveal clinically and biologically relevant partition of the patient cohort.* In this project, we apply a consensus clustering algorithm called Molecular Regularized Consensus Patient Stratification (MRCPS) that previously have been successfully used for predicting outcomes for triple negative breast cancers.

In this work, we integrate MRCPS with gene co-expression network mining to identify combinations of co-expressed gene modules with clinical information that can predict NB patient outcomes, especially the ones that were previously considered "high-risks". The integrated workflow is shown in Fig 1. Since both RNA-seq and gene expression microarray data are available for these NB patients, we take advantage of both types of transcriptomics data.

The sheer large number of features (genes, probesets, etc) in the transcriptomic data poses a challenge on the downstream data integration as well as the statistical power for detecting representative gene expression features. To reduce the data dimensionality and improve statistical power of detection, we first identified densely connected co-expression modules in microarray and RNA-seq data. We used lmQCM (local maximum Quasi-Clique Merger) algorithm to mine co-expressed gene modules and summarized each module into a "eigengene" using the protocol described in [4]. This approach not only substantially improves statistical power, but also greatly reduces the data dimensionality and distills the molecular features of the important biological processes, functions or genetic variants, which facilitates the down-

stream integration with other data types and the interpretation of the results. Next in the workflow, we combined the module eigengenes and clinical data together, and applied a computational method to regularize the clinical classification using molecular weight matrix as prescribed by the affinity of sample in the molecular features space that was defined according to molecular subtypes and estimate density-based models. In the cases that the initial clustering led to a stratification of the patients without significant difference in survival times (i.e., log-rank test p value less than 0.05), we switch to patient similarity matrix based graph method to integrate with clinical information. In this paper, we incorporated these two methods to obtain weighted patient similarity matrix from transcriptomic data and integrate it with categorical clinical attributes from the same patient cohort and pursued a consensus clustering of the cohort.

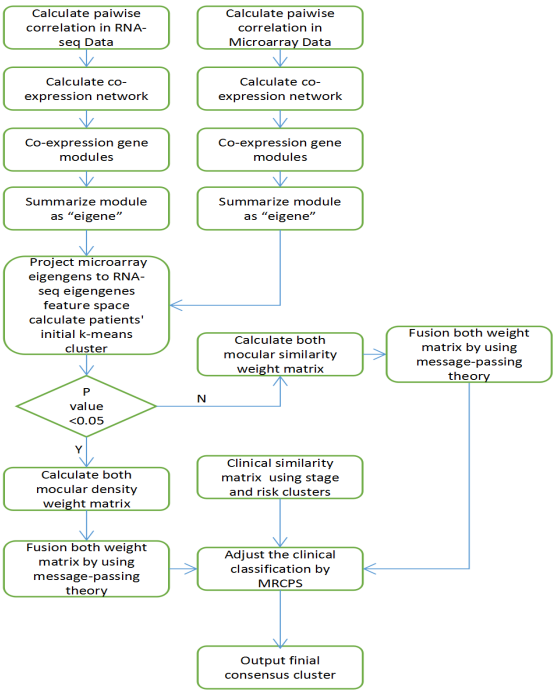


Figure 1: Integration of molecular features with clinically-defined patient stratification workflow

2 Methods

2.1 Dataset and preprocessing

Transcriptome datasets are obtained from Neuroblastoma Data Integration Challenge of CAMDA 2017 for 498 pediatric patients with known clinical endpoints. The data including RNA-seq for 60,788 transcripts and Agilent microarray data for 45,198 probesets. We identified 9,583 genes whose profiles are present in both RNAseq and mircoarray datasets with matched gene symbols for further analysis and data integration.

2.2 Gene co-expression analysis and summarization

We applied our recently developed weighted network mining algorithm lmQCM[4] for gene co-expression module mining. Unlike the popular algorithm WGCNA that utilizes hierarchical clustering therefore does not allow overlaps between clusters, our algorithm takes a greedy mining approach, allowing genes to be shared among multiple gene modules, agreeing with the fact genes often participate in multiple biological processes. In addition, we have shown that lmQCM can find smaller co-expressed gene clusters that are often associated structural mutations such as copy number variations in cancers. The lmQCM algorithm uses four parameters, namely γ,α, t, andβ. Among these parameters, $\gamma$is the most influential as it decides if a new module can be initiated by setting the weight threshold for the first edge of the module as a subnetwork. In our analysis, we transformed the absolute values of the Spearman correlation coefficients between expression profiles of all pairs of genes as weights using a weight-normalization procedure adopted from spectral clustering [4]. β specifies the threshold for overlap ratio between two modules. If the overlap ratio between two modules (defined as the ratio between the size of overlap and the size of the smaller module) is larger than *beta*, the two modules will be merged into a larger one. In practice, we found with γ=0.80, $t = 1$, $\alpha = 1$, and $\beta = 0.4$ yielded gene modules with reasonable sizes (less than 500 genes). Specifically, it identified 38 co-expressed gene clusters of mircoarray and 24 co-expressed gene clusters of RNA.

2.3 Molecular Regularized Consensus Patient Stratification

We previously developed a mathematical formulation for integrative clustering of multiple-modal data. Specifically, we introduced a consensus clustering method MRCPS based on an optimization process with regularization [5]. We built two kinds of MRCPS using molecular density weight matrix and the molecular similarity weight matrix respectively, to ensure the effectiveness of our consensus cluster method. We adopted the maximum mutual information to statistically evaluate the patient cluster number k[5]. This workflow is flexible. we can change the patient similarity matrix based on the molecular data according to the data distribution.

2.3.1 Patient Similarity Matrix based on Molecular Data

Cluster density function[6]: Based on the molecular features, a clustering algorithm such as K-means can be applied thus each patient i is clustered in its molecular subgroup. Then, we can define a cluster density function f of this sample. A classic choice of the density function is the Gaussian Kernel density function[7]: where K_h is a Gaussian Kernel function with parameter h and N_i is the number of patients in the same cluster with features $x_i \in \Re^p$.

$$f(i) = \frac{1}{h^p N_i} \sum_{j=1}^{N_i} K_h(x_i - x_j)$$
$$= \frac{1}{N_i (2\pi h^2)^{\frac{p}{2}}} \sum_{j=1}^{N_i} exp\left(-\frac{\|x_i - x_j\|}{2h^2}\right)$$

For patients i and j, we have defined density estimators $f(i)$ and $f(j)$ respectively as the density of clusters they belong to. This density function denotes the ''molecular affinity'' of the cluster sample which contains i. We can assign weight $W(i,j) = f(i) \times f(j)$,if $i \neq j$ and i, j are in the same cluster while $W(i,j) = 0$,if $i \neq j$ and i, j in the different cluster. Finally, $W(i,j) = 1$,if $i = j$.

2.3.2 Patient Similarity Matrix Using Similarity Network Fusion

In the cases that the initial clustering using the above matrix led to a stratification of the patients without significant difference in survival times (i.e., log-rank test p value *leq* 0.05), we can define the weight matrix using a nonlinear method based on message-passing theory. This method has been previously adopted in Similarity Net Fusion [8] to integrate data from multiple sources. Specifically, for a patient similarity network, edge weights are represented by an $n \times n$ similarity matrix W with $W(i,j)$ indicating The similarity between patients i and j. $W(i,j)$ is generated by applying a scaled exponential similarity kernel on the Euclidean distance $d(x_i,x_j)$ between the patient features x_i and x_j [8]:

$$W(i,j) = exp\left(-\frac{d^2(x_i,x_j)}{\mu\varepsilon_{i,j}}\right),$$

where μ is an empirical hyperarameter set at 0.3 and $\varepsilon_{i,j}$ is defined as

$$\varepsilon_{i,j} = \frac{mean(d(x_i,N_i) + mean(d(x_j,N_j) + d(x_i,x_j)}{3}.$$

Through the above method we obtain the molecular weight matrices for microarray and RNA-seq datasets respectively. Then they are merged using the message-passing method mentioned above.

2.4 Categorical distance metric

In order to use the distance matrix from transcriptomic data to refine the patient clusters defined by the clinical attributes, we first need to define a distance metric for the clinical similarity between a pair of samples. The categorical distance metric between two clinical clusters C^l,C^k is

$$dist(C^l, C^k) = \sum_{i<j} \left[S^l_{ij} - S^k_{ij}\right]^2,$$

where S^l_{ij}=1 if the patient samples i and j are in the same cluster, and otherwise is 0.

Next, we take the weight matrix generated from the molecular data to adjust the clinical clusters. We weighed each pair of genes' similarity $S_{i,j}$ with the fused Molecular Weight Matrix W for every i and j. The underlying rationale is that, if two patient samples i and j are in a cluster of poor molecular clustering result, similarity between them should be low. Thus, a lower weight is given to leverage the high clinical similarity $S_{i,j}$. Given a set of L as the clinical partitions, we can optimize the following cost function to find the optimal partition of patients:

$S^\star = \frac{1}{L} \arg\min_{S^\star} \sum_{i=1}^{L} \sum_{i<j} w_{i,j} \left[s^l_{ij} - s^\star_{ij}\right]^2$

3 Results

3.1 Predicting prognosis for the entire patient cohort

To evaluate the prognostic performance of our method, we compared our results (ie, Kaplan-Meier curves and log-rank test between survival times of patients in different clusters) with clinical features (ie, clinical stage or risk level) alone (Figure 3(a) and 3(b)). Specifically, we tested integration of the two types of transcriptomics data with clinical stage, risk level, as well as both clinical stage and risk level using two approaches. The first is to use the original MRCPS algorithm to calculate the patient similarity matrix as described in Section 2.3.1 (Figure 4). The second approach is to use the message passing approach as described in Section 2.3.2 (Figure 5). In addition, for each approach, we also compared the results with those obtained using transcriptomics data alone. But since the MRCPS algorithm was designed to integrate both transcriptomics and clinical data, we used the similarity network fusion algorithm for transcriptomics data (Figures 4(a) and 5(a)).

As shown in Figure 3, the clinical factors such as stage and risk level can effectively stratify patients into groups with significantly different survival times. Specifically, when the factor divides patients into low-risk and high-risk groups while the pathology information separates patients into five stages (1,2,3,4s,4). The results are shown in Fig.3(a) and Fig.3(b) (log-rank $p = 9.21e-30$ and log-rank $p = 3.88e-37$). Clinical risk are better than clinical stage in prognosis, as all of the low risk patients survived.

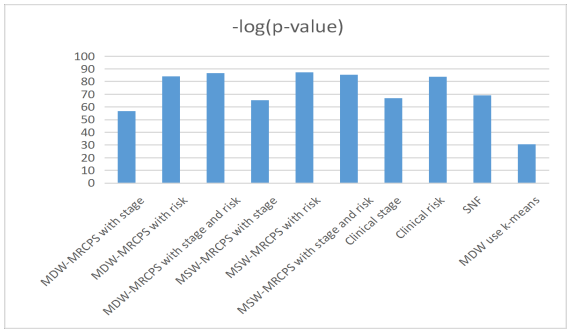
The prognostic prediction results of using transcriptomic data alone are shown in Figures 4(a) and 5(a). While the patients are well separately, the prediction is inferior than using clinical factors, suggesting that integration clinical factors may improve the prediction. And the integrative analysis results confirmed this notion as shown in the rest of the figures. Both molecular weight matrices of MRCPS generate better prognosis than clinical prognosis and independent molecular cluster, as shown in Figure 4(c) and Figure 5(d), as the survival curves show log-rank p-values of 1.16e-38 and 2.08e-38, respectively. As clustering results show, MRCPS makes full use of clinical information, and has superior capability to separate patient populations with different outcomes. Specifically, MRCPS using both molecular weight matrix identified a subtype that has significantly poorer survival rate of less than 40%.

One observation is that the predictions shown in Figures 4(c) and 5(d) are even better than using the risk levels. Since all patients in the low-risk group survived, this observation suggest that the transcriptomic data may also improve of the prediction for high-risk patients alone and we next focus on the high-risk group.

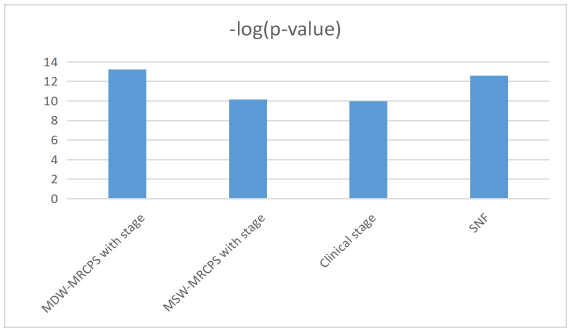
Predicting prognosis for high-risk patients

After applying MRCPS to NB data, we discovered that with both Molecular weight matrix the high risk patients assigned to different subgroups and a higher risk subgroup can be identified. We want to know if the clustering results by using our method is better than using disease stage

in high risk patients. So we applied our method for high risk patients. These results, together with the induced clinical partitions obtained by disease stage are shown in Figure 6(c). Although clinical features separated patients of low risk very well, it does not further stratify the high risk group enough. MRCPS clustered high risk patients into subgroups, as shown in Figure 6(b) and 6(d). The clustering result of using molecular similarity weight matrix is worse than using the clinical stage, for molecular similar weight matrix using spectral clustering, we found that k=2 is the best cluster result according to maximum mutual information, the result is shown in Figure 6(a), it is difficult to reconcile with the five clinical stages. The result of molecular density weight matrix of MRCPS showed better results of, retaining clinical stage and a more accurate classification according to log-rank test. In particular, patients of stage IV in clinical were divided into two groups, where the 84% patients was in the group 4 and group 5 of the new clustering result. Group 5 has a worst prognosis, with the survival rate reduced to less than 40% in first 50 months.The -log(p-value) of These result -log(p-value) shown in Fig.2



(a) 498 NB patients



(b) 239 NB patients in High risk group

Figure 2: Compare -log(p-value) of predict the survival outcomes between multiple method in all NB patients and high risk patients

4 Discussion and Conclusion

In this paper, we developed a workflow to integrate the transcriptomic data and clinical data of NB patients. While the currently used clinical factors can predict patient outcome well, our results showed that our workflow has a superior performance for the entire cohort when both types of transcriptomic data are integrated.
In addition, we found that the previously identified "high-risk" group of patients can be further stratified into multiple groups with significantly different prognosis. While a subgroup of patients with extremely poor survival in early months was identified, a group of high-risk patients actually demonstrated good prognosis (Figure 6(d)).
In the meanwhile, we applied and tested two kinds of molecular affinity matrix, and the proposed MRCPS of molecular density weight matrix method can better stratify patients into repeatable and clinically relevant subtypes as demonstrated by the results. This method can be also extended to the integration of other kinds of genomic features like copy number, somatic mutations, SNP, and pathological information as well as for other cancer types.

References

[1] Stefano Moretti Sara Stigliani, Simona Coco. High Genomic Instability Predicts Survival in Metastatic High-Risk Neuroblastoma. *Neoplasia*, 14(9):823–832, 2012.

[2] Jim E. Griffin Bernard J. de la Cruz Richard S. Savage, Zoubin Ghahramani and David L. Wild1. Discovering transcriptional modules by Bayesian data integration. *BIOINFORMATICS*, 26:158–167, 2010.

[3] Joydeep Ghosh Alexander Strehl. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

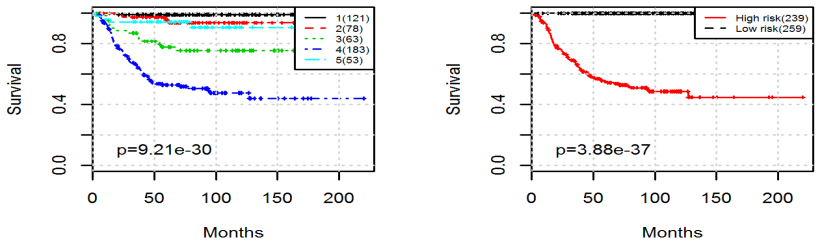
[4] Jie Zhang, Kewei Lu, Yang Xiang, Muhtadi Islam, Shweta Kotian, Zeina Kais, Cindy Lee, Mansi Arora, Hui wen Liu, Jeffrey D. Parvin, and Kun Huang. Weighted Frequent Gene Co-expression Network Mining to Identify Genes Involved in Genome Stability. *PLoS Computational Biology*, 8(8), 2012.

[5] Chao Wang, Raghu Machiraju, and Kun Huang. Breast cancer patient stratification using a molecular regularized consensus clustering method. 67(3):304–312, 2014.

[6] J. Sander A. Zimek Wiley Interdiscipl. Rev. H.P.Kriegel, P. Kröger. Density-based clustering. *Data mining and knowlegdge discovery*, 1:304–312, 2011.

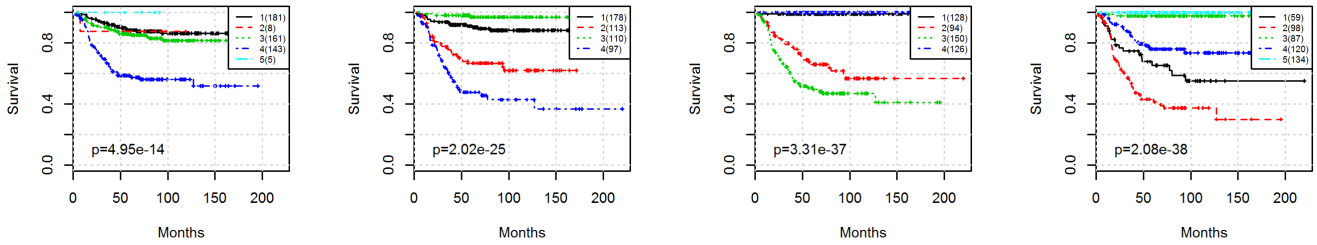
[7] Theory Probabil V.A. Epanechnikov. Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability Its Applications*, 14(1):153–158, 2014.

[8] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.



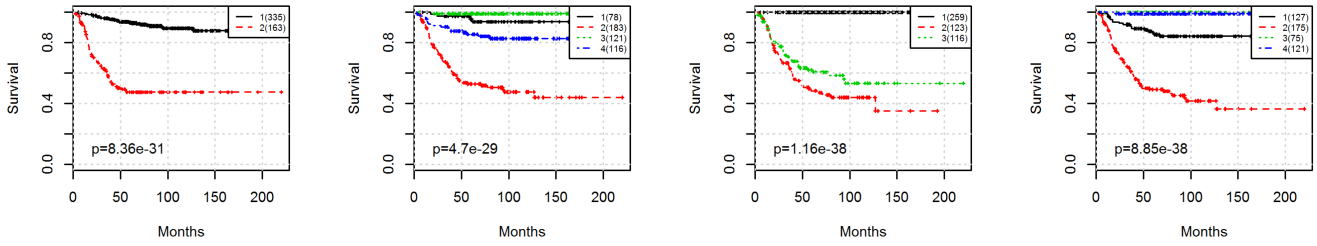
(a) Clinical stage (b) Clinical risk

Figure 3: Clinical diagnose information predict the survival outcomes



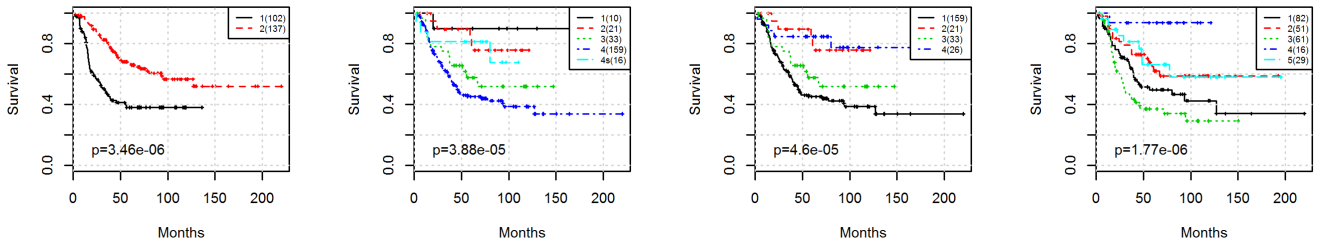
(a) Molecular density weight matrix with kmean cluster (b) Integrate with clinical stage (c) Integrate with clinical risk (d) Integrate with clinical stage and risk

Figure 4: MRCPS of Molecular density weight matrix predict the survival outcomes



(a) Similarity network fusion (b) Integrate with clinical stage (c) Integrate with clinical risk (d) Integrate with clinical stage and risk

Figure 5: MRCPS of Molecular similar weight matrix predict the survival outcomes



(a) Similarity network fusion (b) Clinical stage (c) MRCPS of Molecular similar weight matrix (d) MRCPS of Molecular density weight matrix

Figure 6: Compare predict the survival outcomes between multiple method in high risk patients